Structure
Opinions

Collecting Butterflies and the Protein Structure Initiative: The Right Questions?

Thomas A. Steitz[1,2,3,*]
[1]Departments of Molecular Biophysics & Biochemistry
[2]Department of Chemistry
[3]Howard Hughes Medical Institute
Yale University, New Haven, CT 06520-8114, USA
*Correspondence: peggy.eatherton@yale.edu
DOI 10.1016/j.str.2007.11.005

I still remember the day in the spring of 1963 when I saw the first atomic structure of a protein in three dimensions. It changed my life. Max Perutz was presenting the Dunham Lectures of the Harvard Medical School in a large gymnasium filled with well over 1000 attentive listeners in the audience. Part way through everyone put on a pair of red-green stereo glasses, and a slide of the atomic structure of myoglobin was put up on a screen whose size was some three times Max's height. After a short time for adjusting the superposition of the two images, the molecule popped into three dimensions over Max's head, and a loud "whoo" emanated from the entire audience. Then he began to explain its fascinating features that no one had ever seen before. I decided then and there that this was how I wanted to explore and understand biological questions.

The question is (as it always is) in science: "What is the question"? It was clear to me from the beginning that the question of major significance to biology was not just "What is the structure of protein X"?, which did continue to be of some general interest for a number of years, but "How can we understand the structural basis of a protein's function in a biological process"? In 1965 the second protein structure was determined, that of lysozyme and its complex with a substrate, thereby providing the first insights into the structural basis of catalysis. It was clear. Structure could inform on biological function!

Over the years I have developed a structural biology theology: in order for structural studies to provide understanding of a biological process, one must know the structures of the entire assembly that executes that process, captured at each step in the process. The structures of the pieces of a clock do not inform on how a clock works and even the structure of the whole clock—in a single state—does not show how it functions. Understanding how a horse runs requires a series of photos capturing the horse in each stage of the process so that they can be put together to make a movie, as was done in that first movie.

I often ask, "What if we knew (as we someday shall) the structure of every protein (and RNA) encoded in the human genome and could pin them on a wall like a butterfly collection? Would that tell us how they work?" We could compare their common and different features and derive their evolutionary relationships, but such taxonomy does not tell us how a butterfly flies, which I consider a more interesting question.

I surmise that a large fraction of the protein and RNA products of the human genome constitute components of large macromolecular assemblies: the clock problem. Cloning, expressing, purifying, and crystallizing separately all 14 subunits of the eukaryotic RNA polymerase would be difficult, if not impossible. Most of the subunits would aggregate due to their exposed hydrophobic surfaces. But even if all of their separate structures could be obtained, of what conceivable value would that be? These structures would not inform at all on how RNA polymerase works. Indeed, neither does the structure of the whole apo-RNA polymerase in the absence of its appropriate substrate complexes.

So what are the strengths and weaknesses of the Protein Structure Initiative (PSI)? The magnitude of the effort has enabled, both directly and indirectly, the development of better technologies that are essential to structure determination, as has funding for structural biology in other formats: cloning, expression, purification, crystallization, and synchrotron facilities have all been improved over the past decade. Methods development is essential to any field and has been ongoing in X-ray crystallography since the early days of Bragg. It is perhaps important to ask what methods need to be developed, and what is the best way to support their development. The goal of solving all protein structures through the PSI may not be the most efficient way to develop new methods. Perhaps a better way would be to provide grants whose goal is the development of new methods in structural biology that are available to individuals or groups without their being part of a larger consortium or their being able to apply for them only once in five years.

One of the major goals of structural genomics suggested by the NIH has been to determine the structures of all the protein folds (RNA not to be included). Why is this an important goal, even if there is agreement on what constitutes a separate fold? Homology modeling a homologous sequence onto a known structure, even if one knows what it does and how it works, is not usually sufficient to understand the function of the unknown structure unless the two proteins are close homologs. Knowing that a protein has a β-barrel structure is of limited use in understanding its function or how it works. It simply is not obvious to me how many structures of this type will have as large an impact on biology

as the way in which structural biology has been executed historically and is done presently outside the PSI.

How should the impact of the PSI compared with the "business as usual approach" be evaluated? One way would be to determine the total number of citations to research emanating from PSI compared with other structural biology research done using the same amount of money during the same period. An attempt to accomplish this on a very limited scale was published last year (Chandonia and Brenner, 2007). Citations to 104 randomly selected manuscripts were compared with 104 random structural biology manuscripts, and the median number of citations was 4 and 11, respectively. I suspect that an evaluation based on the total number of structures solved would show a more significant difference. Comparisons were done on a limited scale, as well, of the costs per structure solved and the numbers of novel folds discovered, and the differences were not large.

What course of action would increase the impact of structural biology on the field of biology the most in the future? In my view two of the most important goals should be to provide research support for young structural biologists, who will of course be the future of the field, and to develop methods for obtaining the structures of large assemblies that are not abundant. The $60 million per year spent on PSI could provide some 200 RO1 grants. I suggest that committing some of this money to young investigators would produce important structural insights into biology as well as future structural biologists. The creative visions of the future for this adventure in structural biology will come from these young investigators, not from the road map designers.

Several approaches may be needed to allow advances in the structural studies of large assemblies at atomic resolution. First of all, the material needs to be made in quantities of 10 s of mg. This will require the development of multigene expression systems, particularly in eukaryotic cells. Currently, small numbers of proteins can be simultaneously expressed, but expression of 10 to several dozen different proteins at a time is needed. Expression and purification of individual proteins leads to insoluble proteins and degradation. A second challenge facing the community for studying these assemblies and the other "high hanging fruit," particularly in starting labs, is the length of time required to successfully meet the challenges presented by many significant structural problems. In my experience, a decade or more is frequently required to achieve success, so that grants with 4 year time-lines present a challenge. I crystallized *E. coli* recA protein in 1979 and we published its structure in 1992. Our structural studies of T7 RNA polymerase were started in the early 1980s and the first paper was published in 1998, and additional examples abound. Ten years and about 15 person years elapsed between our publication of a presynaptic complex of resolvase with DNA and its synaptic complex with a cleaved DNA intermediate. We have been working for more than 10 years to obtain suitably diffracting crystals of the 70S ribosome complexed with translation factors EFG or EFTu•aatRNA and have yet to be successful. Such problems confront all structural biology labs, of course, but this is where the significant insight into the structural basis of biological functions lies—not in the structures of unliganded domains of proteins.

In summary, I think it is indeed wonderful that a large amount of money is being spent in the field of structural biology, and the PSI goals have certainly been more significant and more successfully accomplished than the crystallization of proteins in space. However, I feel that even greater benefits to the biological community at large would be achieved by redirection of a significant fraction of the funds to: development of both X-ray and molecular biological methods, long-term studies of large complexes, and increased funding available for the younger half of the structural biology community.

REFERENCE

Chandonia, J.-M., and Brenner, S.E. (2007). The impact of structural genomics: expectation and outcomes. Science *311*, 347–351.

